The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance

MERIT-Infonomics Research Memorandum series

The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance

Gerald Silverberg & Bart Verspagen

2004-021

MERIT – Maastricht Economic Research Institute on Innovation and Technology

PO Box 616
6200 MD Maastricht
The Netherlands
T: +31 43 3883875
F: +31 43 3884905

http://www.merit.unimaas.nl
e-mail:secr-merit@merit.unimaas.nl

International Institute of Infonomics

c/o Maastricht University
PO Box 616
6200 MD Maastricht
The Netherlands
T: +31 43 388 3875
F: +31 45 388 4905

http://www.infonomics.nl
e-mail: secr@infonomics.nl

The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance

Gerald Silverberg*

and

Bart Verspagen**

August 2004

Abstract

This paper focuses on the analysis of size distributions of innovations, which are known to be highly skewed. We use patent citations as one indicator of innovation significance, constructing two large datasets from the European and US Patent Offices at a high level of aggregation, and the Trajtenberg (1990) dataset on CT scanners at a very low one. We also study self-assessed reports of patented innovation values using two very recent patent valuation datasets from the Netherlands and the UK, as well as a small dataset of patent license revenues of Harvard University. Statistical methods are applied to analyse the properties of the empirical size distributions, where we put special emphasis on testing for the existence of 'heavy tails', i.e., whether or not the probability of very large innovations declines more slowly than exponentially. While overall the distributions appear to resemble a lognormal, we argue that the tails are indeed fat. We invoke some recent results from extreme value statistics and apply the Hill (1975) estimator with data-driven cut-offs to determine the tail index for the right tails of all datasets except the NL and UK patent valuations. On these latter datasets we use a maximum likelihood estimator for grouped data to estimate the Pareto exponent for varying definitions of the right tail. We find significantly and consistently lower tail estimates for the returns data than the citation data (around 0.7 vs. 3-5). The EPO and US patent citation tail indices are roughly constant over time (although the US one does grow somewhat in the last periods) but the latter estimates are significantly lower than the former. The heaviness of the tails, particularly as measured by financial indices, we argue, has significant implications for technology policy and growth theory, since the second and possibly even the first moments of these distributions may not exist.

JEL Codes: C16, O31, O33
Keywords: returns to invention, patent citations, extreme-value statistics, skewed distributions, heavy tails.

---

* MERIT, Maastricht University, P.O. Box 616, NL-6200 MD Maastricht, The Netherlands, Tel. +31-43-3883868, Fax +31-43-3216518. Email: gerald.silverberg@merit.unimaas.nl.

** ECIS, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands, Tel. +31-40-2475613, Fax +31-40-2474646. Email: b.verspagen@tm.tue.nl.

Life is not an illogicality; yet it is a trap for logicians. It looks just a little more mathematical and regular than it is; its exactitude is obvious, but its inexactitude is hidden; its wildness lies in wait.
G.K. Chesterton

1. Introduction

Innovations are created in a somewhat mysterious process, but they are not all created equal. Some few innovations seem to have major implications, often opening up whole new areas of scientific and technological activity, while others are quickly forgotten and perhaps never even implemented. This has been clearly demonstrated repeatedly on the basis of citations data for both patents and scientific publications. Other data on the financial returns to innovation and R&D also demonstrate a similar (in fact, even more extreme) skewness of the distributions. As early as the 1960s, the presence of skewness and some of its implications had already been recognized (e.g. in Kuznets 1962 and Scherer 1965).

While the extreme skewness of these distributions is now uncontested, the so-called heaviness or fatness of the right tail of the distribution turns out to be a statistical question somewhat distinct from that of the shape of the overall distribution, and also has wide-ranging implications for our understanding of the innovation process. If the tails are Pareto distributed (that is, resemble a power law of the form $x^{-\alpha}$, rather than an exponential like a normal or lognormal), then much more of the activity will be concentrated very far from the 'typical' values than would otherwise be the case. And the moments of the distribution of order > α will cease to exist, including in the extreme case ($\alpha<1$) the mean value itself. Thus, to understand the 'riskiness' of innovative activity, it is necessary to estimate the 'fatness' of the distribution. Lack of existence of the variance ($\alpha<2$) alone makes many of the traditional methods of risk analysis and econometrics inapplicable.

The issue of the fatness of the tail is not quite the same as that of the shape of the whole distribution and its extreme skewness. Models that provide a better overall goodness of fit to the entire distribution may seriously underestimate the tail, for which empirical data will only be very sparse and often rejected a priori as outliers. The tail (suitably defined) will only represent a comparatively small part of the probability mass but a large part of the overall impact (in terms of total citations, total returns, etc.) since it continues on so far to the right. Thus an appropriate estimate of the characteristics of the tail distribution is of utmost importance in these cases, independently of the 'best' fit to the overall distribution, which may look quite different. A distribution that satisfies a global goodness of fit criterion may fit the tail very poorly (and vice versa).

Fortunately, extreme value statistics offers us a canonical classification of the possible distributions of the largest observations sampled from an independent (or weakly dependent) and identically distributed process, and some general methods for estimating important statistical characteristics. While these methods have been used in the natural sciences and financial economics for some time, to our knowledge they have not yet been applied to innovation data.[1] The purpose of this paper is to do precisely this.

[1] While Harhoff, Scherer and Vopel (2003) purports to be about the tail of the patent value distribution, these authors use 'tail' only in the sense of observations above a rather low and arbitrary threshold

2. Statistics of Innovation Size Distributions

In this paper we will work with two types of indicators of innovation 'size' or 'significance': patent citations and monetary values. Citations have increasingly become one of the main indicators of scientific and technological significance.[2] With the advent of cheap computing power, they have become relatively easy to compile from electronic databases of patents and scientific publication indexes. It is in the nature of such publications to cite previous relevant work, although there may be incentives (or simply ignorance) to bias this activity. The primary function of patent citations is a legal one, i.e., indicating which parts of the described knowledge can and cannot be claimed by the patent. In the European patent system these citations are for the most part added by the patent examiner, while in the US system the applicants themselves add most of the citations. Trajtenberg (1990) argued that forward citations of a patent (citations by subsequent patents of a given patent) were a good indicator of the economic value of the invention in the restricted class of CT medical scanners. A number of subsequent studies have confirmed the value of forward citations (as well as other variables such as backward citations) with respect to various measures of patent value (as inferred for example from patent renewal rates, self-assessment, financial market values, etc.).[3] Regardless of the economic value of a patent we can regard citations in a purely scientometric sense as self-generated indicators of the technological significance of an invention and of the relations of influence or similarity (subject of course to various caveats). We begin with Trajtenberg's original data and then go on to construct annual cohorts of citations from the entire European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) databases.

We first examine raw 'Pareto' plots of three empirical datasets. These consist of right cumulative distributions or the number of observations with a value greater than or equal to a given amount, plotted on a double log scale. The first is compiled from Trajtenberg's (1990) original dataset on CT scanner patent citations (Figure 1). A true Pareto or power law distribution would be linear. While we do observe a slight curvature, the linearity over practically the entire range is remarkable. The rightmost or most extreme value (corresponding to a patent cited 73 times) actually lies above any regression line and might normally be regarded as an outlier. The second (Figure 2) is compiled from EPO statistics for citations until 1999 to all patents with priority date in 1989. This period is long enough to ensure that most citations a patent will receive in its lifetime are actually being captured. The dataset consists of 33,499 cited patents and 80,928 citations, with the most cited patent being cited 63 times. The tail does appear to be rather linear and extensive. Figure 3 depicts the US patent citations data in this form for 1989, representing 50,687 cited patents and 321,385 citations. The most cited patent is cited 212 times.

(DM23,000) and not in the sense of extreme-value theory. As we shall see, the Pareto tail in the latter sense only begins at much higher quantiles. While their maximum-likelihood estimator is appropriate for a true tail (and we employ a version of it with our grouped data), other work in this field has failed to draw on up-to-date statistical theory to estimate the tail index and discuss goodness-of-fit issues.

[2] For an overview of recent work with patent citations see the contributions in Jaffe and Trajtenberg (2002).

[3] See Hall, Jaffe and Trajtenberg (2000), Harhoff, Scherer and Vopel (2003), Harhoff, Narin, Scherer and Vopel (1999), and Jaffe, Trajtenberg and Fogarty (2000) for some recent findings

Our second measure of patent size will be based on monetary value.[4] Figure 4 presents the Pareto plot of a small dataset for the financial returns from licensing fees to patents granted to Harvard University.[5] Here, only the tail shows any linearity, but with something of a high outlier for the second-largest observation. Our final datasets derive from a recent survey of self-evaluations of patent values by the patent inventors in a number of European countries.[6] In contrast to the previous datasets, the data here consists of grouped rather than point observations. Figure 5 and Figure 6 show the results for the Netherlands (NL) and the United Kingdom (UK), respectively. Because of the group nature of the observations, these datasets will be analysed with different methods than the others.

While there can be no dispute that all of these datasets (and many others examined in the literature, see Scherer 1998) are all highly skewed, there has been considerable uncertainty regarding whether they are better represented by a fat-tailed distribution such as a Pareto or a highly skewed but 'medium'-tailed distribution such as the lognormal. To some extent this boils down to the question whether we are interested in the overall shape of the distribution or the behaviour of the extreme values in the tails. A goodness-of-fit criterion will address the former question but not necessarily the latter. Scherer (1998) argues in many cases for the greater plausibility of the lognormal in terms of the overall fit. Harhoff, Scherer and Vopel (2003) maintain this assertion even in an investigation of the behaviour of the distribution tails using maximum likelihood methods.

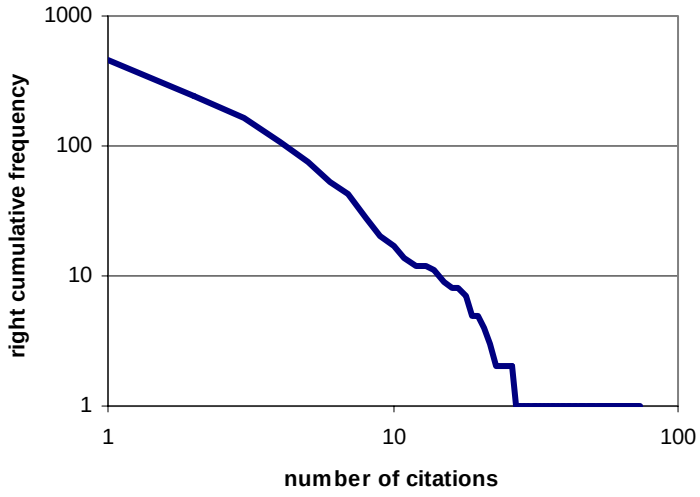


Figure 1. Pareto plot of Trajtenberg patent citation data.

[4] For an investigation of the relationship between citation counts and monetary values see e.g. Trajtenberg (1990), Betrán (2003).
[5] This dataset was kindly provided to us by Frederic M. Scherer.
[6] See Nesta et al. (2004) for a description of the data-gathering techniques and survey questionnaire. We draw the UK data directly from this source, while the Dutch data are from the part of the survey implemented by ECIS, Eindhoven University of Technology.

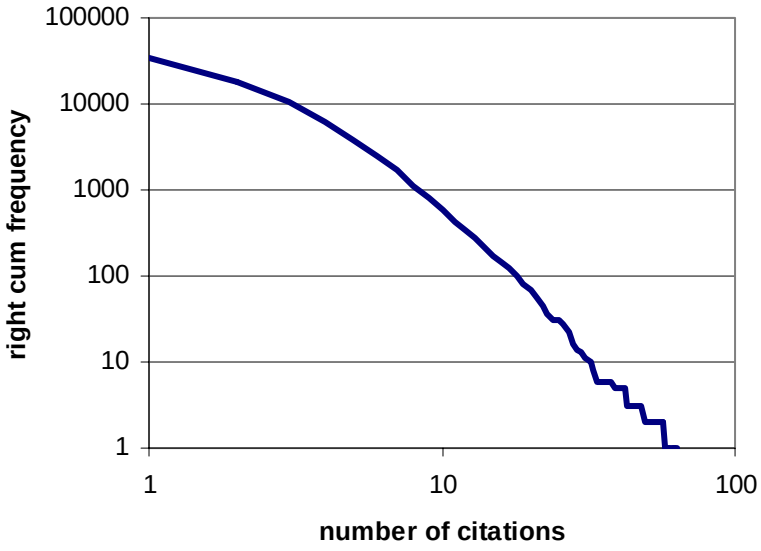


Figure 2. Pareto plot of EPO 1989 patent citation data.

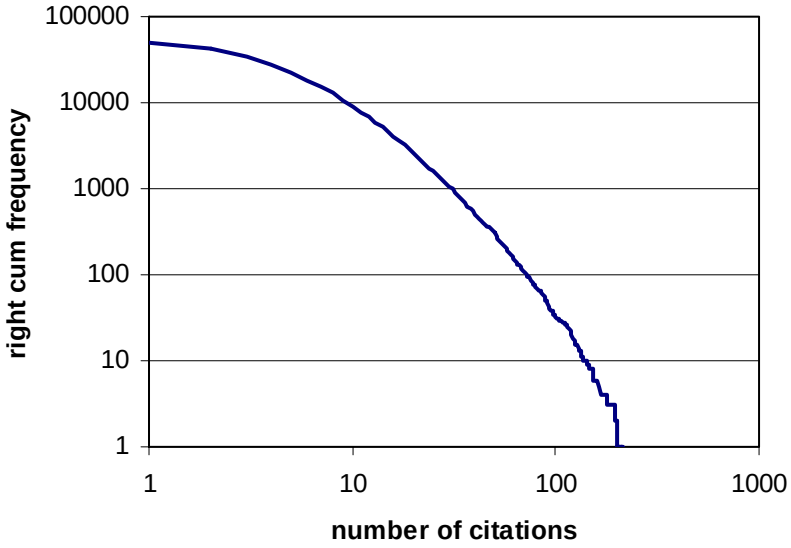


Figure 3. Pareto plot of USPTO 1989 patent citation cohort.

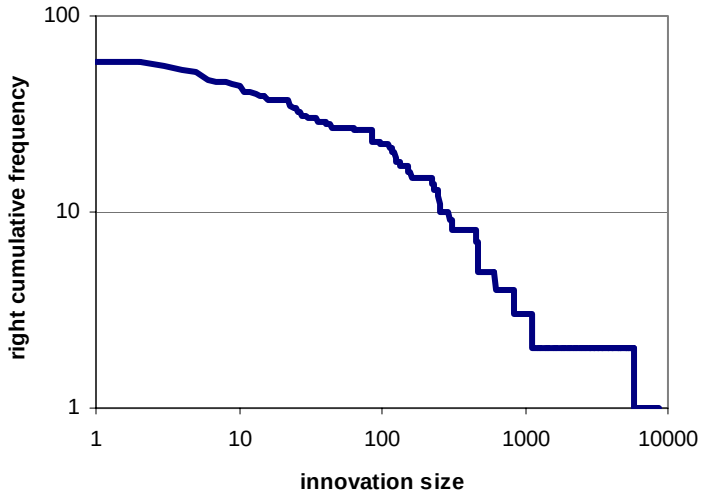


Figure 4. Pareto plot of Harvard patent returns data.

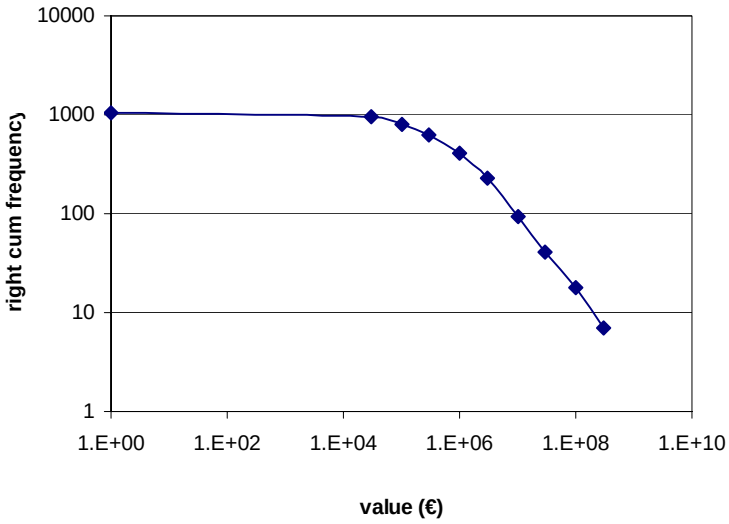


Figure 5. Pareto plot of NL patent valuation survey.

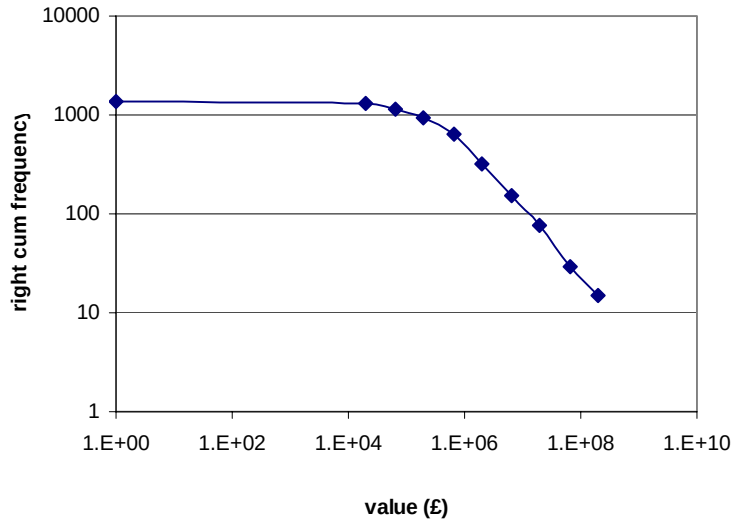


Figure 6. Pareto plot of UK patent valuation survey.

To address the issue of lognormal vs. Pareto, we present p-p plots for the respective distributions in Figures 7-10. It is apparent that over the entire range of data the lognormal provides a much better fit. However, since the Pareto parameter is estimated in these plots using the entire dataset and not just the tail, these Pareto fits will not be optimal for the tail. Since so much of the impact of innovations is contained in the rightmost tail, we will employ more sophisticated methods based on extreme-value theory to scrutinize it more closely. As we shall see, Pareto distributions fitted properly to the appropriately defined tail segments capture the tail behaviour more accurately than the lognormal, despite the latter's overall superiority in terms of aggregate goodness of fit.

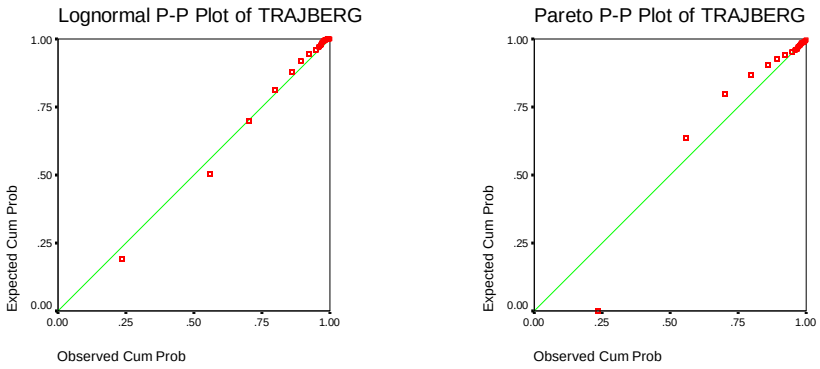


Figure 7. Lognormal (left) and Pareto (right) p-p plots for Trajtenberg patent citation data.

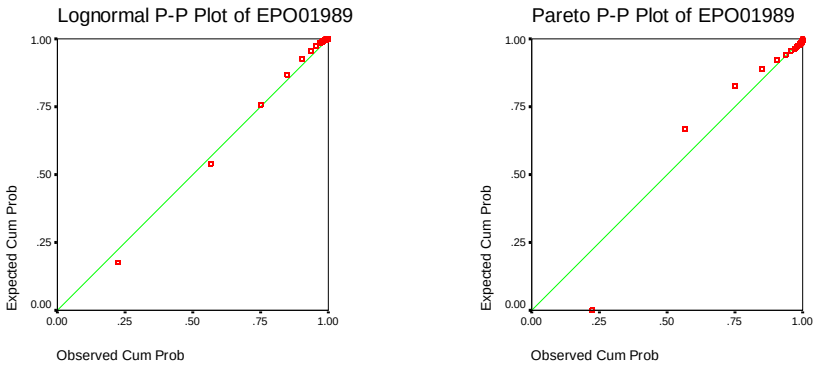


Figure 8. Lognormal (left) and Pareto (right) p-p plots for EPO 1989 citation data.

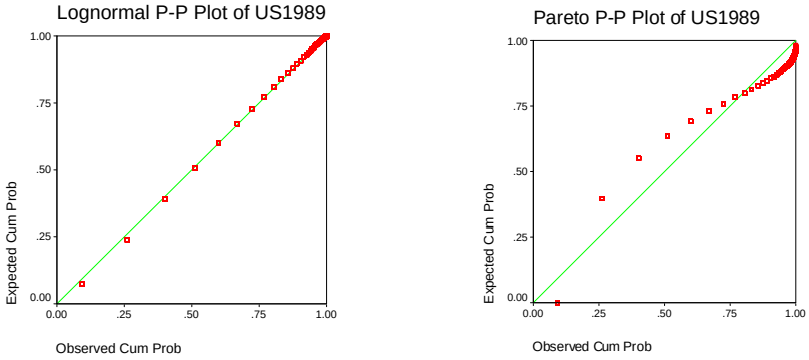


Figure 9. Lognormal (left) and Pareto (right) p-p plots of US 1989 citation data.

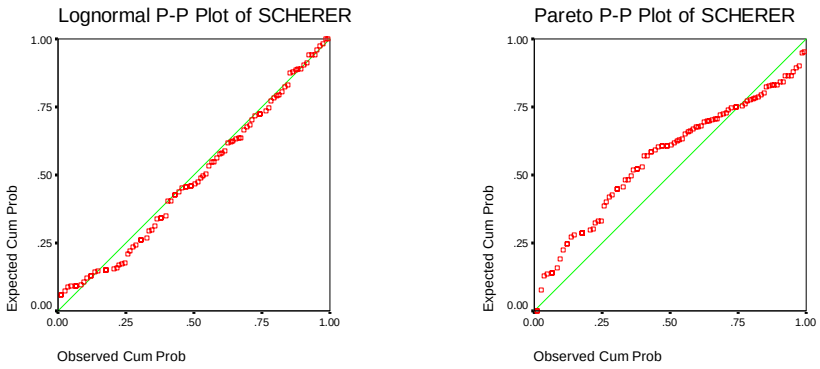
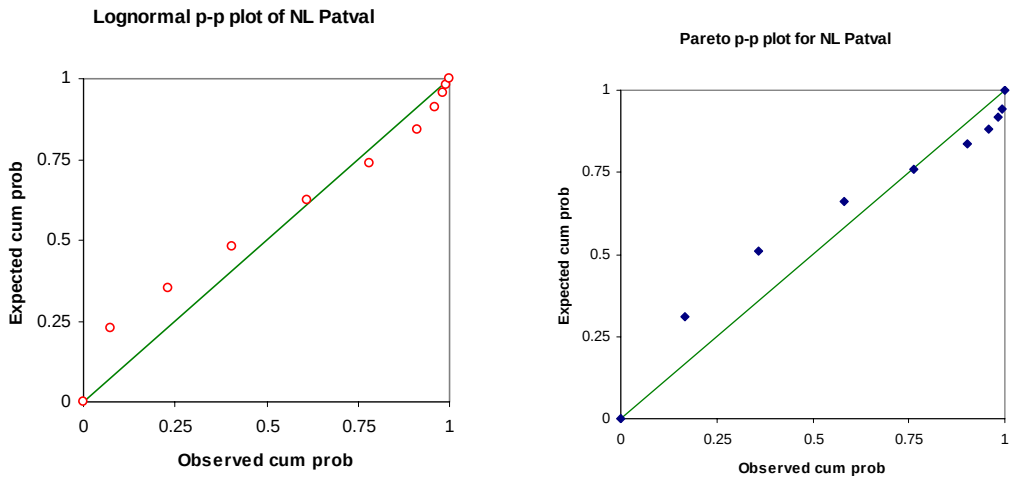


Figure 10. Lognormal (left) and Pareto (right) p-p plot for Harvard patent returns data.



Fig

Figure 11. Lognormal (left) and Pareto (right, with threshold at €30,000) p-p plots for NL Patval dataset.

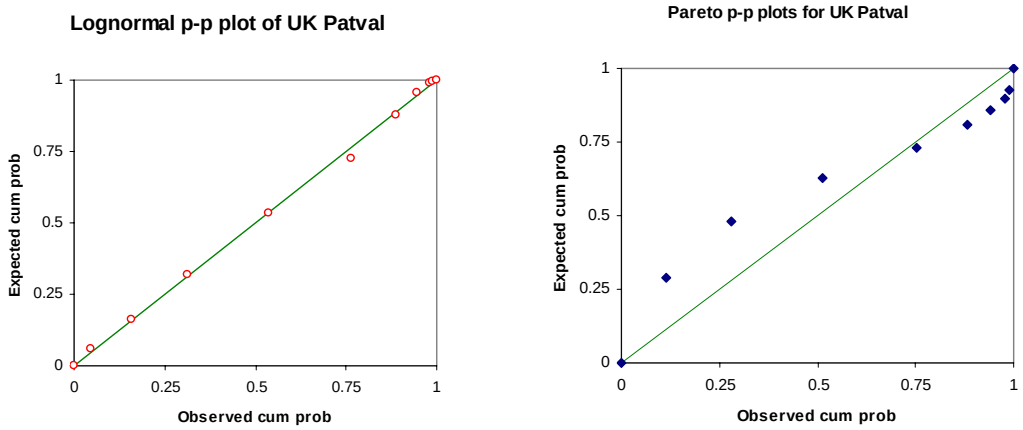


Figure 12. Lognormal (left) and Pareto (right, with threshold at £19,500) plots for UK Patval dataset.

3. The Hill Estimator and Extreme Value Statistics

To understand the behaviour of the tails of our distributions we can draw on important results from extreme value statistics. According to this work (see Embrechts, Kluppelberg and Mikosh 1997, Coles 2001, Reiss and Thomas 2001, Resnick 2004), the extreme values of iid observations of an distribution will in general be in the domain of attractive of one of three limiting distributions. These correspond to the heavy or fat-tailed case (Pareto, stable distributions, Student *t*), short-tailed case (e.g., uniform), and medium-tailed one (normal, exponential distributions). It can be shown that a Generalised Pareto Distribution (GPD) captures the tail behaviour for all three cases, with the fat-tailed case corresponding asymptotically to the ordinary Pareto distribution. A simple maximum likelihood estimator for the exponent parameter α of the GPD distribution was introduced by Hill (1975). Placing the n observations X_i in

descending order and denoting the resulting rank-order statistics by $X_{[i]}$, $X_{[1]} \geq X_{[2]} \geq \ldots \geq X_{[n]}$, the Hill estimator is defined as follows:

$$H(k,n) = \frac{1}{k}\sum_{i=1}^{k}(\ln X_{[i]} - \ln X_{[k+1]}).$$

Plotting this estimator against k for small values of k (compared to n) will indicate if it converges to some value, which will then be an estimate for the downward slope of the double-log rank-order plots (so-called Zipf plots), or the inverse of the exponent α (= $1/H$) of the estimated Pareto-Levy distribution:

$$N = \kappa X^{-\alpha},$$

where X is the value of an observation, N is the number of observations with value X or larger and κ and α are positive parameters. It can be shown that the expression $(H-h)k^{1/2}$, where H is the estimate and h is the true value, is asymptotically normal with mean zero and variance h^2.

A crucial problem in using the Hill estimator for a distribution that is only Pareto in the tails to estimate a Generalized Pareto Distribution is to determine a cut-off value for the tail. The further to the left the cut-off, the more data is used in the estimation, but the more the behaviour may deviate from the 'Pareto-ness' of the tail. This can be seen in a diagram of the Hill estimator as a function of k, the number of data points of the rank order statistics entering into the calculation. The trade-off between increasing the sample size (and thus reducing the variance) and increasing the bias of the estimate can make estimating α from the Hill diagram alone highly subjective. A number of modifications of the Hill estimator have been proposed that somewhat reduce its volatility. These include moment and QQ estimators and a smoothed Hill (see Resnick 2004). There are also good grounds for looking for a criterion for cut-off determination driven by the data themselves.

A number of methods have recently been developed that hinge primarily on minimizing the mean squared error of the Hill estimator as a function of k. A good summary of recent work in this direction can be found in Lux (2001), who also applies these methods to a large financial dataset. Due to computational limitations we will only report here the results for the method of Drees and Kaufmann (1998).

4. Results for Patent Citation and R&D Returns Data

Figure 13 to Figure 16 plot H as a function of k for our point-observation datasets. The estimated value is reasonably stable over certain ranges of k but nevertheless varies considerably over the whole range.[7] The vertical lines show the cut-off values k^* for the tails determined by the method of Drees and Kaufmann and used in the estimation of α. With the exception of the Harvard patent returns data, there is no evidence for a value of α below 1. And the Harvard returns case, as we have seen, otherwise most closely resembles a lognormal, so this may be some kind of statistical artefact. Or it may be evidence of the greater impact of the right tail in returns than in citations. Thus the innovation process, at least as far as patent citations can be used as a proxy for significance, does not seem to fall into the pathological cases of infinite mean. Higher moments, however, may not be finite.

[7] The saw tooth nature of the citation data curves is due to the discrete-valued nature of these datasets and the fact that many patents share the same number of citations for low citation values.

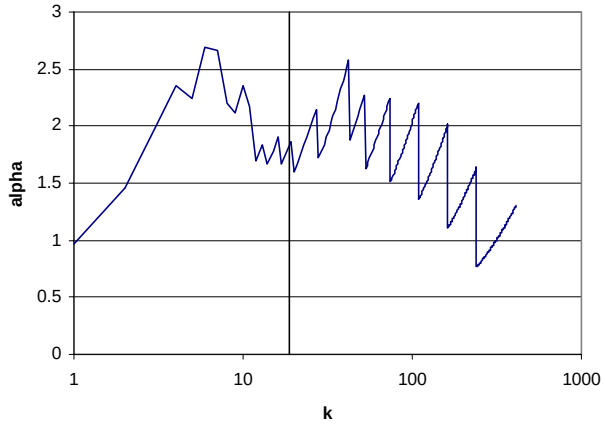


Figure 13. Hill estimator of α for Trajtenberg patent citation data (log scale). Vertical line shows the Drees/Kaufmann cut-off value.

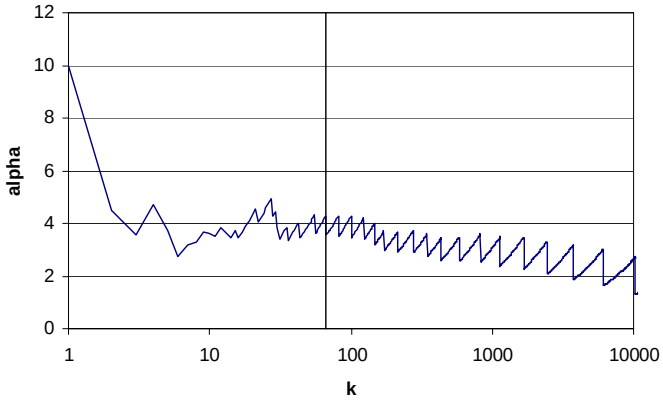


Figure 14. Hill estimate of α for EPO 1989 citation data (log scale).

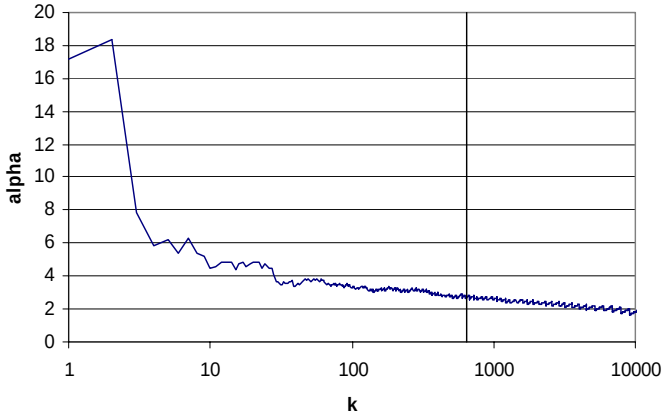


Figure 15. Hill estimate of α for US 1989 citation data (log scale).

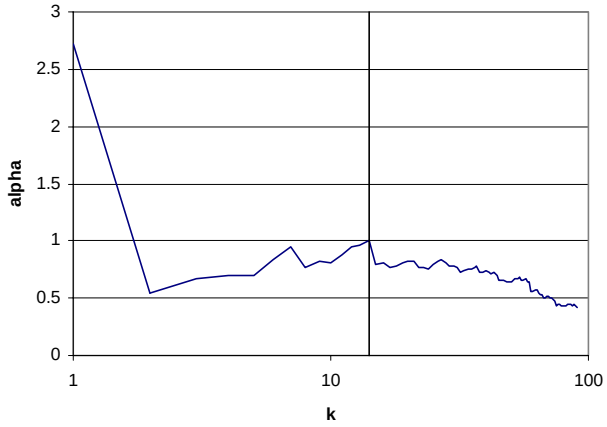


Figure 16. Hill estimate of α for Harvard patent returns data (log scale).

The estimates of α determined by the method of Drees and Kaufmann are summarized in Table I. Using a different method due to Danielsson and de Vries (1997) that we were only able to apply to the shorter datasets due to computation limitations, we obtained α estimates of 0.812 and 2.332 for the Harvard and Trajtenberg datasets, respectively, with cutoffs at 19 and 38. The estimates for the shorter datasets should be taken with a grain of salt, although comparably few observations are actually used even for the immensely larger EPO and US datasets.

Table I. Estimates of α by the Drees and Kaufmann method

Dataset	α	confidence interval		threshold	k*	n
Harvard	1.010	0.480	1.537	$230K	14	100
Trajtenberg	1.864	1.026	2.701	9	19	456
EPO1989	3.542	2.694	4.390	20	67	33499
US 1989	2.718	2.509	2.927	36	650	50687

For the EPO and USPTO patent citations datasets we have computed an estimate of α using the Drees and Kaufmann method for each cohort of cited patents. The point estimates and confidence intervals are plotted in Figure 17. The estimates are remarkably consistent between years within the same dataset, though the estimated α for the EPO 1996 cohort appears to be something of an outlier. The EPO estimates are also almost always higher (except for four years), and just about significantly so, than the US ones (EPO mean 3.91, sample σ 0.83, US mean 3.15, sample σ 0.39, with the point estimates of one dataset lying near or outside the boundaries of the confidence intervals of the other). The individual Hill plots on which these estimates are based, as well as the results of significance tests using the moment method of the hypothesis that $1/\alpha \neq 0$, as well as alternative estimates based on a quantile method (see Resnick 2004), can be examined on the web[8].

[8] http://www.tm.tue.nl/ecis/bart

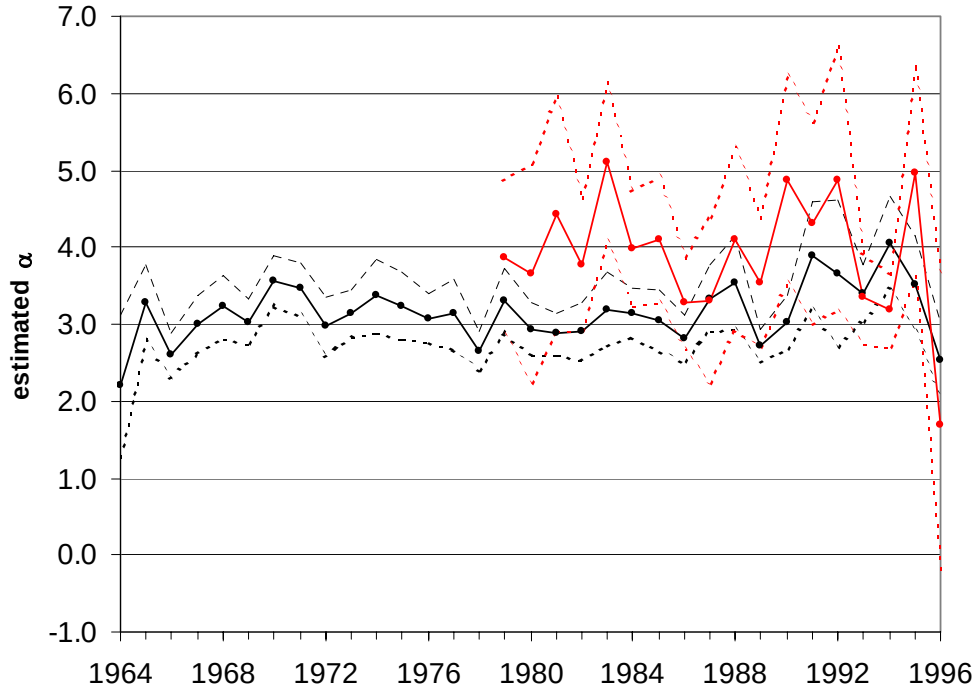


Figure 17. Drees/Kaufmann estimates of α for each cited patent cohort of EPO dataset (red) and USPTO dataset (black), with 95% confidence intervals (dotted lines).

5. Tail Index Estimates for NL and UK Patent Valuation Surveys

Recently, Internet-based and telephone surveys of patent holders have been undertaken in several EU countries to determine the value of patents based on a standardized questionnaire. This has involved asking the patent inventor to state the minimum price he/she would think the patent holder (i.e., usually the inventor's employer) ask a potential competitor interested in buying the patent on the day it was granted. This is very similar to the type of question used in the survey in Harhoff, Scherer, and Vopel 2003. We draw on the results for the Netherlands (NL) and the United Kingdom (UK) in the following. Since the survey asks the respondents to place the value of the patent into one of several intervals, the datasets are not comparable to the previous ones since they represent grouped data rather than point observations.

For the Netherlands the bounds of these intervals are € 30,000 and below, 100,000, 300,000, 1 million, 3 million, 10 million, 30 million, 100 million, and 300 million and above, while for the UK they are £ 19,500 and below, 65,000, 195,000, 650,000, 1,950,000, 6,500,000, 19,500,000, 65 million, and 195 million and above. The datasets consist of 967 (NL) and 1302 (UK) observations. We can compute the log likelihood function of a truncated Pareto distribution on the right tail of these observations from some interval bound L_c onwards using the distribution function of the truncated Pareto, $\text{Prob}\{x \geq y\} = (y/L_c)^{-\alpha}$, as follows:

$$\log L(\alpha, L_c) = \sum_{i=c}^{m-1} n_i \log[(L_i / L_c)^{-\alpha} - (L_{i+1} / L_c)^{-\alpha}] + n_m \log(L_m / L_c)^{-\alpha},$$

where L_i is the lower bound of the ith interval (and the upper bound of the i-1th interval), m is the index of the last, unbounded interval, α is the Pareto parameter to be estimated, and n_i is the number of observations in the ith interval $[L_i, L_{i+1})$ (see Falk, Hüsler and Reiss 1994, p. 140, Scherer, Harhoff and Vopel 2003, Technical Appendix). Maximizing the log likelihood over α yields the maximum likelihood estimate of the Pareto parameter. As with the Hill estimator, this will be a function of the threshold L_c.

Table III. Estimates of α by the ML method on grouped data

Dataset	α	confidence interval		threshold	k*	n
NL Patval	0.743	0.639	0.847	€3M	228	1046
UK Patval	0.632	0.578	0.686	£650K	633	1368

In Figure 18 and Figure 19 we plot the results with the associated confidence intervals for the NL and UK datasets, stepping through values of L_c, calculated using the maximum likelihood routine from the *tsp* package. Table II documents the results numerically. We see that for the NL data a plateau exists for thresholds between € 3 million and 100 million, with an estimated value of α of 0.732-0.743, providing plausible evidence for a GPD tail. A similar plateau exists for the UK data for thresholds between £ 650,000 and 19.5 million, yielding estimates in the range 0.632-0.742. Clearly, estimating the Pareto parameter over the entire dataset, or from very low thresholds such as in the case of Scherer, Harhoff and Vopel (2003) (who employ DM 23,000 as their tail threshold), provides a highly biased estimator of the tail index and a poor goodness of fit to the relevant range of data.[9] Our analysis indicates that the 'tail' in the sense of extreme value theory actually starts around € 3 million (NL dataset) and £ 650,000 (UK dataset), corresponding to the largest 228 (or 24%) and 633 (or 49%) of the observations, respectively. What stands out here even more so than in the case of the Harvard data is that the estimates are significantly below one, indicating the most pathological of tail behaviours: both infinite mean and infinite variance of the distributions.

[9] The fact that the rightmost point estimates deviate from the plateau values, in either direction, undoubtedly reflects the high variance of the estimator at the extreme tail of the dataset, where only observations in the two highest intervals are used.

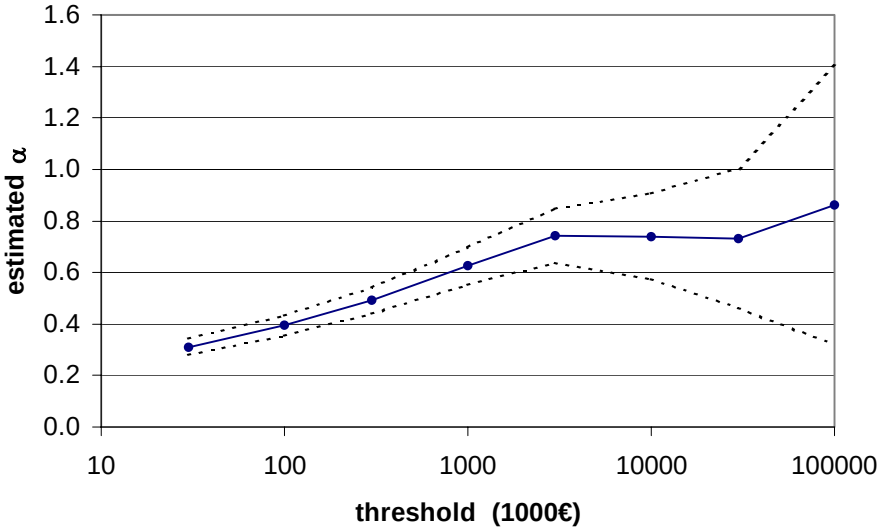


Figure 18. Maximum likelihood estimate of α with 95% confidence intervals for Dutch Patval dataset at various thresholds (log scale).

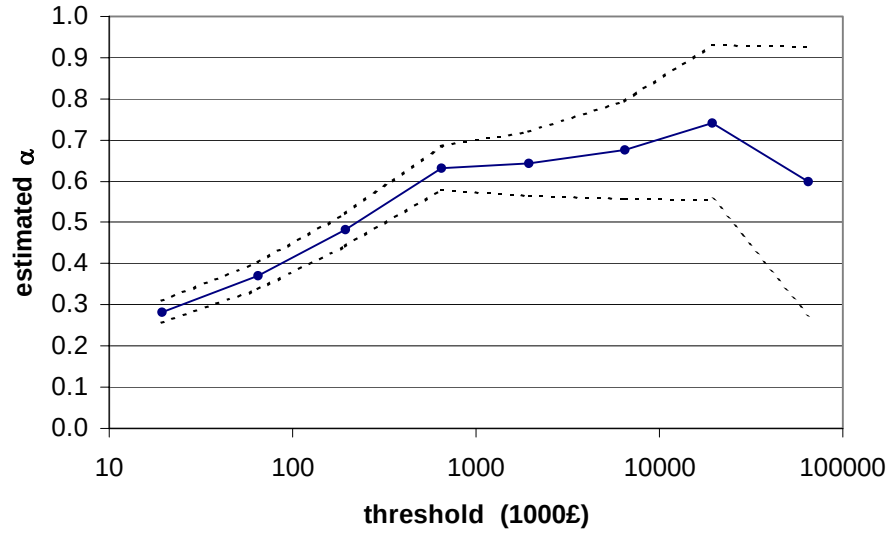


Figure 19. Maximum likelihood estimate of α with 95% confidence intervals for UK Patval dataset at various thresholds (log scale).

The goodness of fit of the distributions can be evaluated by examining p-p plots based on truncated datasets beginning at the thresholds determined by the plateaus in the estimated αs. If the dataset is truncated at x^*, then the cumulative lognormal distribution $F_{ln}(x)$ must be replaced by $(F_{ln}(x)-F_{ln}(x^*))/(1-F_{ln}(x^*))$ in the calculation of the diagrams. The lognormal can also be fitted to the truncated dataset instead of to the entire dataset by making use of the truncated log-likelihood function (see Harhoff, Scherer and Vopel 2003, Technical Appendix):

$$\log L(\theta) = \sum_{i=c}^{m-1} n_i \log[(F_{\ln}(L_{i+1},\theta) - F_{\ln}(L_i,\theta)) / F_{\ln}(L_c,\theta))] + n_m \log[(1 - F_{\ln}(L_m,\theta)) / F_{\ln}(L_c,\theta)],$$

where $F_{ln}(x,\theta)$ is the cumulative lognormal probability distribution, θ is the vector of

scale and shape parameters, L_c is the truncation boundary, and m is the index of the last, open-ended interval. The p-p plots for the tails starting at € 3 million (NL) and £650,000 (UK) are shown in Figure 20 and Figure 21. It is clear that the Pareto provides a better fit to the tail when this is appropriately defined and used in the estimation of the tail index itself, than the lognormal fitted to the entire dataset. Of course, fitting a lognormal specifically to the tail region improves the fit of this distribution to the truncated dataset, even if it markedly alters the parameter estimates[10] and completely sacrifices the lognormal's close fit to the entire dataset. Nevertheless, the Pareto still appears to be a superior fit to the tails of the two datasets, although the differences are small and probably not statistically significant. Of course, there is no theoretical justification for fitting a lognormal to the extreme tail of such a dataset, since the lognormal is not one of the canonical functional forms of the Generalized Pareto Distribution (in fact, the exponential distribution corresponds to the tail of the lognormal class of distributions).

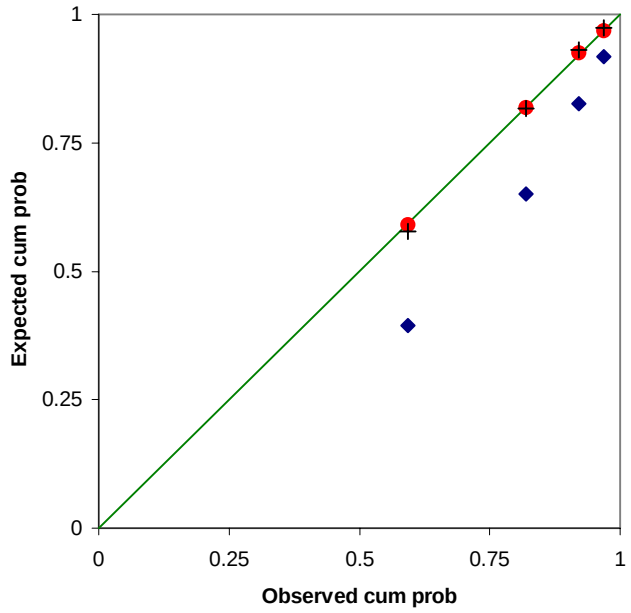


Figure 20. p-p plots for NL Patval dataset left-truncated at € 3 million, lognormal fitted to entire dataset (blue diamonds), lognormal fitted to tail (+), Pareto with α = 0.743 fitted to tail (red discs).

[10] Thus the mean and variance of the underlying normal distribution shifts from (12.77, 3.33) to (0, 4.85) for the NL data and from (13.20, 2.12) to (0, 5.08) for the UK data.

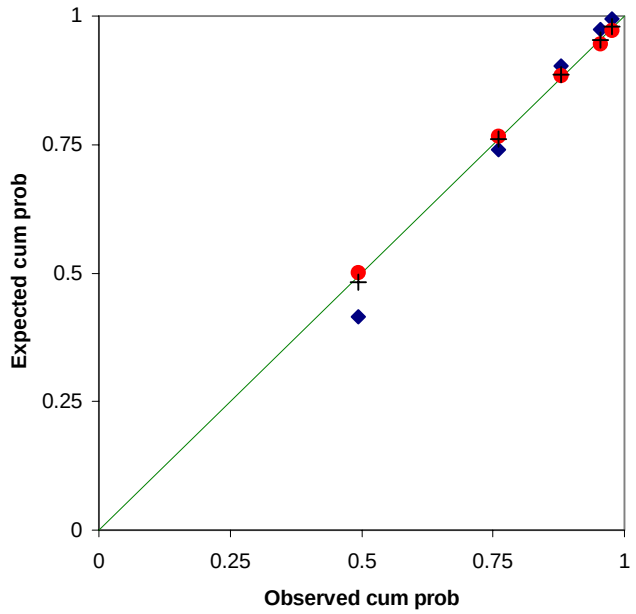


Figure 21. p-p plots for UK Patval dataset left-truncated at £650,000, lognormal fitted to entire dataset (blue diamonds), lognormal fitted to tail (+), Pareto with $\alpha = 0.632$ fitted to tail (red discs).

6. Implications of Skewed and Fat-Tailed Innovation Distributions for Technology Policy and Management

It has been recognized in the literature for some time that the highly skewed and possibly fat-tailed distribution of returns from R&D projects poses a tricky problem for the management of innovation (see, e.g., Scherer and Harhoff, 2000). Since only a small percentage of all projects yield a positive return (on the order of 20-30%), and more than half of all returns (by some estimates more than 90%) are generated by the top 10% of all projects, a diversified portfolio is necessary to achieve reasonable technological change. Yet the riskiness of the portfolio does not decline rapidly with its size, while there are arguments that its mean actually increases (equivalent to increasing returns to the size of the portfolio, cf. Sornette 2002). The latter argument is a possible explanation for the tendency of pharmaceuticals to merge into every larger units, since the firms are dependent on only a handful of drugs, of the many investigated, turning out to be blockbusters (the same may be argued for Hollywood film studios, see e.g. Vany and Walls 2004).

While the overall shape and skewness of these distributions is itself of great interest, the possibility of fat tails has a disproportionate influence on the riskiness of innovation portfolios. Thus, as in the financial arena, it is necessary to examine the tails more closely before appropriate management policy can be formulated.

A presupposition of our statistical methods and of technology policy as portfolio management has been that the realizations are independent. However, one could argue that many of the intermediate and minor innovations are actually consequences of major paradigmatic innovations that open up whole new fields or

methodologies. In that case perhaps a rather different statistical approach would be called for that takes into account the persistence of innovation activity across size classes and intertemporally.

7. Conclusions and Directions for Future Research

We have examined three empirical datasets on the size distribution of innovations based on financial returns, and three based on citations. All display the well-known property of extreme skewness. Although the overall shape of the distributions appears to be more lognormal than Pareto, we have argued that the tail behaviour needs to be analysed from the perspective of extreme value statistics to be dealt with correctly.

This approach argues that the statistical behaviour of the distribution's tail under a wide range of assumptions can only take one of three canonical forms. The highly skewed, fat-tailed case is the one of particular interest in this connection, where the Generalized Pareto Distribution becomes relevant. We have applied the Hill estimator to the point-observation datasets and a straightforward maximum likelihood estimator to the grouped data. The results from applying the Hill estimator with data-determined tail cut-offs do indeed indicate that we are dealing with Pareto-like tails, while the existence of a tail 'plateau' for the two grouped datasets seems plausible. The returns datasets all yield an estimate in the critical region at or below $\alpha=1$. In contrast, the large citations datasets seem to lie in a more 'stable' region with α between 3 and 4, with the medium-sized CT citations dataset lying between the two. However, datasets with higher estimated values of α are difficult to differentiate from lognormal-distributed data, as Monte Carlo experiments show.

Acknowledgements

We wish to thank Thomas Lux for putting his Gauss scripts and some of his expertise at our disposal, and Frederic M. Scherer for providing the Harvard dataset.

References

Betrán, F.L., 2003, "Pricing Patents through Citations", University of Rochester: working paper.

Coles, S., 2001, *An Introduction to Statistical Modeling of Extreme Values,* Berlin: Springer-Verlag.

Danielsson, J. and de Vries, C. G., 1997, "Tail Index and Quantile Estimation with Very High Frequency Data", *Journal of Empirical Finance,* **4:** 241-257.

Drees, H. and Kaufmann, E., 1998, "Selecting the Optimal Sample Fraction in Univariate Extreme Value Estimation", *Stochastic Processes and their Applications,* **75:** 149-172.

Embrechts, P., Kluppelberg, C. P. and Mikosh, T., 1997, *Modelling Extremal Events,* Berlin: Springer-Verlag.

Falk, M., Hüsler, J. and Reiss, R.D., 1994, *Laws of Small Numbers: Extremes and Rare Events,* Basel/Boston/Berlin: Birkhäuser.

Hall, B. H., Jaffe, A. and Trajtenberg, M., 2000, "Market Value and Patent Citations:

A First Look", Cambridge, MA: NBER Working Paper No. 7741.
Harhoff, D., Narin, F., Scherer, F. M. and Vopel, K., 1999, "Citation frequency and the value of patented inventions", *Review of Economics and Statistics,* **81:** 511-515.
Harhoff, D., Scherer, F.M. and Vopel, K., 2003, "Exploring the Tail of Patented Value Distribution", in Grandstrand, O., (ed.), *The Economics of the Patent System,*
Hill, B. M., 1975, "A Simple General Approach to Inference about the Tails of a Distribution", *The Annals of Statistics,* **3:** 1163-1174.
Jaffe, A.B., Trajtenberg, M. and Fogarty, M.S., 2000, "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees",
Jaffe, A.B. and Trajtenberg, M. (eds), 2002, *Patents, Citations and Innovations: A Window on the Knowledge Economy,* Cambridge MA and London: MIT Press.
Kuznets, S., 1962, "Inventive Activity: Problems of Definition and Measurement", in Nelson, R. R., (eds), *The Rate and Direction of Inventive Activity: Economic and Social Factors,* Princeton: Princeton University Press.
Lux, T., 2001, "The Limiting Extremal Behaviour of Speculative Returns: An Analysis of Intra-Daily Data from the Frankfurt Stock Exchange", *Applied Financial Economics,* **11:** 299-315.
Nesta, L., Crespi, G., Geuna, A., Brusoni, S. and Patel, P., 2004, "The UK Survey on Patent Value (PatVal). Methodological Report", SPRU, University of Brighton, March.
Reiss, R.D. and Thomas, M., 2001, *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology, and Other Fields,* Basel/Boston/Berlin: Birkhäuser.
Resnick, S., 2004, "Modeling Data Networks", in Finkenstaedt, B. and Rootzen, H., (eds), *Extreme Values in Finance, Telecommunications, and the Environment,* London: Chapman & Hall.
Scherer, F. M., 1965, "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions", *American Economic Review,* **55:** 1097-1123.
Scherer, F. M., 1998, "The Size Distribution of Profits from Innovation", *Annales d'Economie et de Statistique,* **49/50:** 495-516.
Scherer, F. M. and Harhoff, D., 2000, "Technology policy for a world of skew-distribution outcomes", *Research Policy,* **29:** 559-566.
Sornette, D., 2002, "Economy of scale in R&D with block-busters", *Quantitative Finance,* **2:** 224-227.
Trajtenberg, M., 1990, "A Penny for your Quotes: Patent Citations and the Value of Innovations", *Rand Journal of Economics,* **21(1):** 172-187.
Vany, A.S. and Walls, W. D., 2004, "Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar", *Journal of Economics Dynamics and Control,* **28:** 1035-1057.